文章类型：研究论文
**DOI**:

# 基于 LSTM 的复杂炼化过程报警预测

韩家祺[1]
（北京化工大学机电工程学院[1]）

**摘要**：近年来，随着我国炼化行业与信息技术的深入融合与飞速发展，复杂炼化系统所产生的数据呈现爆炸性增长。报警系统是一类用于向操作者传递设备异常状态信息的监控系统；一旦设计不合理，设备在异常状态下可能会产生大量的过程报警甚至报警饱和的现象，严重影响操作者的信息处理能力，从而增加各种工业事故的发生概率。报警信息能够对复杂炼化过程给予正向的指导，因此如何从海量的报警日志中挖掘有价值的信息非常重要。深度学习是一种能够自动地从数据中学习和提取特征的方法，不需要人工构建复杂而精确的物理和数学模型，已在数据预测和分类领域得到广泛应用和关注。

**关键词**：炼化过程；报警管理；自然语言处理；深度学习
**中图分类号**：TQ 028.8

# Alarm prediction of complex refining process based on deep learning

**HAN Jiaqi[1], XU Zhinan[2], FAN Mei[1,2], CEN Peilin[2]**
([1] *College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology*, *Beijing* 100029)

**Abstract**: In recent years, with the rapid development of the chemical industry and information combination, the data produced in the chemical refining system presents explosive growth. Alarm system is a kind of transmitting equipment abnormal state information to the operator of the system, but if the design is not reasonable, the equipment under abnormal state process may produce a large number of alarm and alarm saturation phenomenon, the serious influence the operator's information processing ability, thus increasing the probability of all kinds of industrial accidents. Therefore, how to mine useful information from the massive alarm logs is very important, and use the mined information to give positive guidance to the complex refining process. Deep learning is a method that can automatically learn and extract features from data. It does not require manual construction of complex and accurate physical and mathematical models, so it has been widely applied and paid attention to in the field of data prediction and classification.

**Key words**: Refining and chemical process; Alarm management; Natural language processing; Deep learning

## 引　言

报警预测是指利用数据分析和算法，根据历史和实时的信息，预测未来可能发生的危险事件。报警预测可以提高报警管理的效率和精度，使得复杂炼化过程安全性提高。同时可以促进数据共享和交流，增强过程数据的价值。目前报警预测大多基于过程数据，比如数字报警，这类报警只有两个值（0 或者 1），不能够获取有价值的时间序列在一定程度上限制了报警预测方法。而基于深度学习的算法模型的输入必须是数值形式的计算机语言，这对于报警日志文本非常困难，需要把纯文本的形式转换为数值形式的计算机语言才能够利用深度学习工具。同时报警日志文本中隐含了关联报警信息，如何在转化的同时保留这一特征也是一个难点。

为了解决不能够获取有价值的时间序列问题，本文提出了利用知识图谱以及深度学习，提出了一种基于 LSTM（Knowledge Graph-Long Short Term Memory）的复杂炼化过程报警预警方法。传统的时间序列预测在实现非线性建模方面有困难，导致预测精度不高，而深度学习可以很好的解决此问题。

# 1 研究背景

炼化行业是国家发展的重要支柱和动力源泉，在我国的现代化发展进程中发挥着关键作用。然而，炼化生产过程复杂，涉及的危险化学品种类繁多，安全风险较高，一旦设备出现过程故障而未能及时处理，很可能导致系统失稳，从而对人员的生命财产安全造成严重危害。因此，实时监测炼化装置的运行状态，及时发现和预防故障，对提高炼化系统的安全性具有重要意义。报警日志是炼化装置运行过程中产生的重要数据源，记录了温度、压力、流量等参数以及报警事件和操作指令等过程信息。通过对这些信息的分析和挖掘，可以实现对炼化装置的故障诊断和实时状态监测，从而为优化工艺控制和改善安全管理提供支持。

炼化行业自动化水平的日渐提高，尤其是分布式控制系统（Distributed Control System, DCS）、先进过程控制系统（Advanced Process Control, APC）以及数据采集与监视控制系统（Supervisory Control And Data Acquisition, SCADA）的广泛应用，使得复杂炼化过程报警装置的成本和性能得到很大改善。随着计算机控制水平的进步，复杂炼化过程正向着规模化、精细化、智能化和集成化的方向发展，大幅度减少了生产过程中的人工干预，优化了设备的可靠性，进一步提升了炼化设备设施的安全管理水平。同时，复杂炼化的规模化和集成化使得过程数据的体量日益增大，函需借助人工智能技术解决人力处理效率不足的问题。

人工智能的飞速发展，让大数据、深度学习和多模态等高新技术焕发出新的活力，知识图谱作为其中的重要分支，被广泛应用于众多领域。当前人工智能正在经历从感知智能到认知智能的发展阶段，而认知智能的本质就是对知识的获取和应用，知识图谱则可以帮助计算机识别人类知识、组织网络资源，进而用知识赋能各个行业的智能应用，例如搜索引擎、智能问答、推荐系统、自然语言处理和机器翻译等。知识图谱及其知识引擎技术作为为人工智能系统的基础，为人工智能的发展提供了强大的支持和保障。

# 2 基本理论

## 2.1 词嵌入技术

词嵌入技术是一种将自然语言中的词汇映射到低维向量空间的技术，它可以有效地表示词汇的语义和语法信息，以及词汇之间的相似性和关联性。词嵌入技术是自然语言处理领域的基础技术，它可以为各种自然语言处理任务提供有用的特征，比如文本分类、命名实体识别、情感分析、机器翻译、文本生成等。词嵌入技术有很多种方法，比如基于计数的方法、基于预测的方法、基于神经网络的方法等。其中，最具代表性的是 Word2Vec 和 Onehot 两种

方法，它们分别利用了局部上下文和全局统计信息来学习词向量。词嵌入技术是人工智能领域的一个重要研究方向，它有助于提高自然语言处理的效果和效率，也有助于挖掘自然语言中的深层次知识。

Word2vec 是一种基于神经网络的词嵌入模型，它可以使用一个双层的神经网络来从大规模的语料库中学习每个词语的 n 维向量表示（n 是嵌入空间的维度）。这种模型可以有效地利用语料库中词语的局部上下文信息，使得在语料库中具有高频共现关系的词语，在嵌入空间中具有较高的余弦相似度和欧氏距离。这样，词向量就能反映出词语在上下文中的语义和语法特征。Word2vec 有两种主要的训练方法，分别是连续词袋模型（CBOW）和跳字模型（Skip-gram），它们分别以不同的方式利用上下文信息来预测目标词语或者以目标词语来预测上下文信息。二者相比 CBOW 模型耗时更短，但是 Skip-gram 在表达出现频次较少的词语方面表现优异。由于报警日志中的报警序列的出现次数很有可能是几次甚至一次，选用 Skip-gram 模型作为词嵌入模型。Skip-gram 工作原理如图 1 所示，分为输入层（Input layer）、投影层（Projection layer）、输出层（Output layer）。

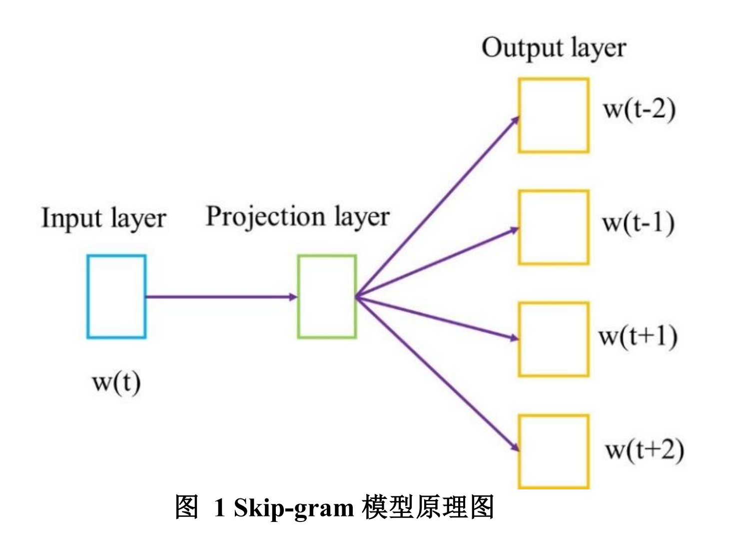



**图 1 Skip-gram 模型原理图**

## 2.2 长短时记忆网络

长短时记忆网络（Long Short-Term Memory, LSTM）是一种特殊的循环神经网络（Recurrent Neural Network, RNN），可有效地解决循环神经网络在处理长序列时遇到的梯度消失和梯度爆炸的问题。目前长短期记忆网络模型是适用最为广泛的循环神经网络模型，与标准循环神经网络模型相比，长短期记忆网络可更好地对长时依赖关系进行表达。如图 2 所示，是一个典型循环神经网络的结构原理图，包含输入层（Input layer）、隐含层（Hidden layer）、

循环层（Cycle layer）、输出层（Output layer），每个时间步接收一个输入，并输出一个输出，同时将自身的隐藏状态传递给下一个时间步。这样，循环神经网络可以利用隐藏状态来存储和利用序列中的历史信息，从而实现对序列的建模。

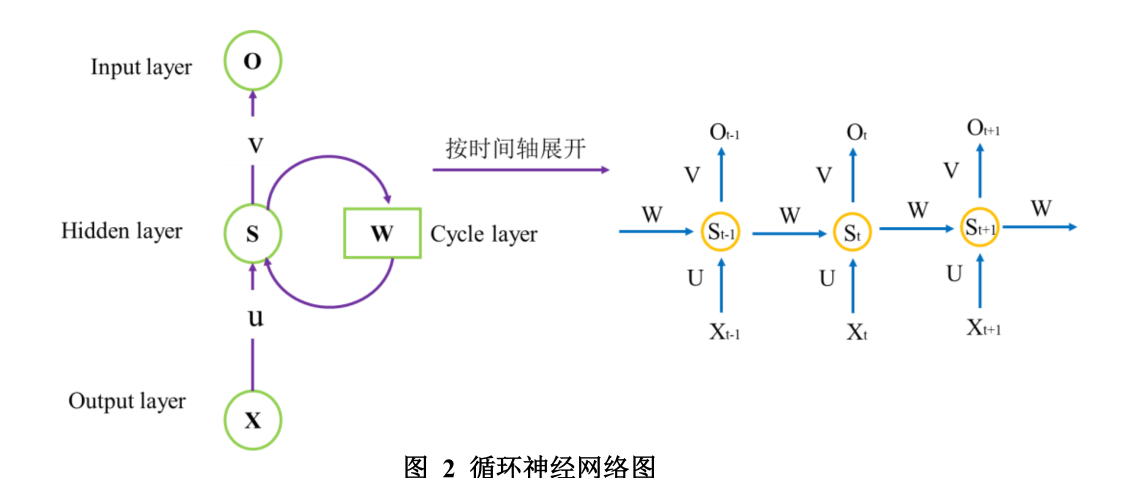



**图 2 循环神经网络图**

长短时记忆网络的核心思想是引入了一个称为记忆单元（Memory cell）的结构，它可以存储和更新长期的信息，以及通过三个门控机制（输入门、遗忘门和输出门）来控制信息的流动。长短期记忆网络可以学习到序列中不同时间步之间的长期依赖关系，从而提高了序列建模的能力。长短期记忆网络是自然语言处理、语音识别、图像生成等领域的一个重要技术，它为各种序列到序列的任务提供了强大的模型，也为后续的深度学习模型提供了灵感和基础。如图 3 所示，是一个典型的长短期记忆网络网络结构图，具有一个输入层、一个输出层及两个在时间维度上展开五步的隐藏层。

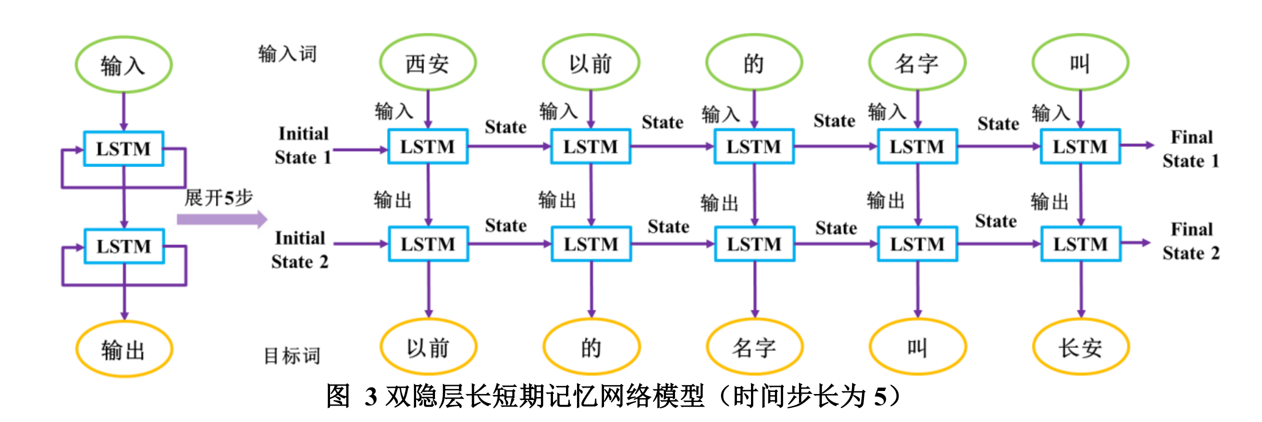



**图 3 双隐层长短期记忆网络模型（时间步长为 5）**

长短期记忆网络一个典型应用就是预测文本的下一个单词，如图 3 所示，对于每个词语来说，与之相邻的词语是我们的期望目标。例如，图 3 中输入“叫”，预测得到的期望输出为“西安”。长短期记忆网络充分利用前面的词语（“西安”、“以前”、“的”、“名字”）的信息

以提高预测结果的准确性。

记忆单元（长短期记忆网络单元）的核心是记忆单元，它可以在整个链条中传递信息，同时受到四个门的控制：遗忘门、输入、输出门和更新门。遗忘门决定了上一个记忆带的哪些信息被留下或抛弃；输入门决定了当前输入的哪些信息被收纳到记忆带中；输出门决定了记忆带的哪些信息被释放为隐藏状态；更新门决定了记忆带的哪些信息被刷新为新的候选值。

此外，许多学者对长短期记忆网络模型进行了深入研究，其在自然语言处理领域成果颇丰如，机器翻译，语音识别及笔迹识别。基于其在自然语言处理上的成功应用（尤其是预测句子中的下一单词），本文将采用长短期记忆网络 模型对过程报警进行预测。

## 3 方法步骤

第一步，数据清洗。数据的输入形式很大程度上决定了模型的处理效果，在第三章知识三元组的基础上将时间属性提取出来，再对原本的数据集进行清洗。将知识三元组中的中文转为英文，利用 NLTK 对英文进行直接分词，获得一个报警序列。一个报警序列的构成为“时间+位号+具体设备+涉及因素+报警等级”（“time + main + specific + factor + level”）。清洗后的数据作为 Skip-gram 模型的输入获得报警词向量。

第二步，构建报警序列词向量。将报警序列输入 Skip-gram 模型得到相应时间序列的词向量。通过设置合适的窗口等参数得到最优的词向量模型。

第三步，LSTM 训练。将得到的词向量模型作为神经网络模型的输入，知识驱动的长短期记忆网络模型（Knowledge Graph-Long Short-Term Memory）由两层组成，第一层是一个循环神经网络层，可以处理输入的时间序列数据，并输出 80 维的向量。这个层使用了 dropout 来防止过拟合，也使用了 bias 来增强记忆能力。第二层是一个全连接层，可以将 80 维的向量映射到一个标量，作为预测值。这个层使用了线性激活函数，以保持输出的连续性。模型使用平均绝对误差和均方误差作为损失函数，使用 Adam 作为优化器。平均绝对误差和均方误差可以衡量预测值和真实值之间的差距，Adam 可以自适应地调整学习率和动量，以加速收敛过程。模型结构可以参考图 3。实现报警预测的步骤如图 4 所示。

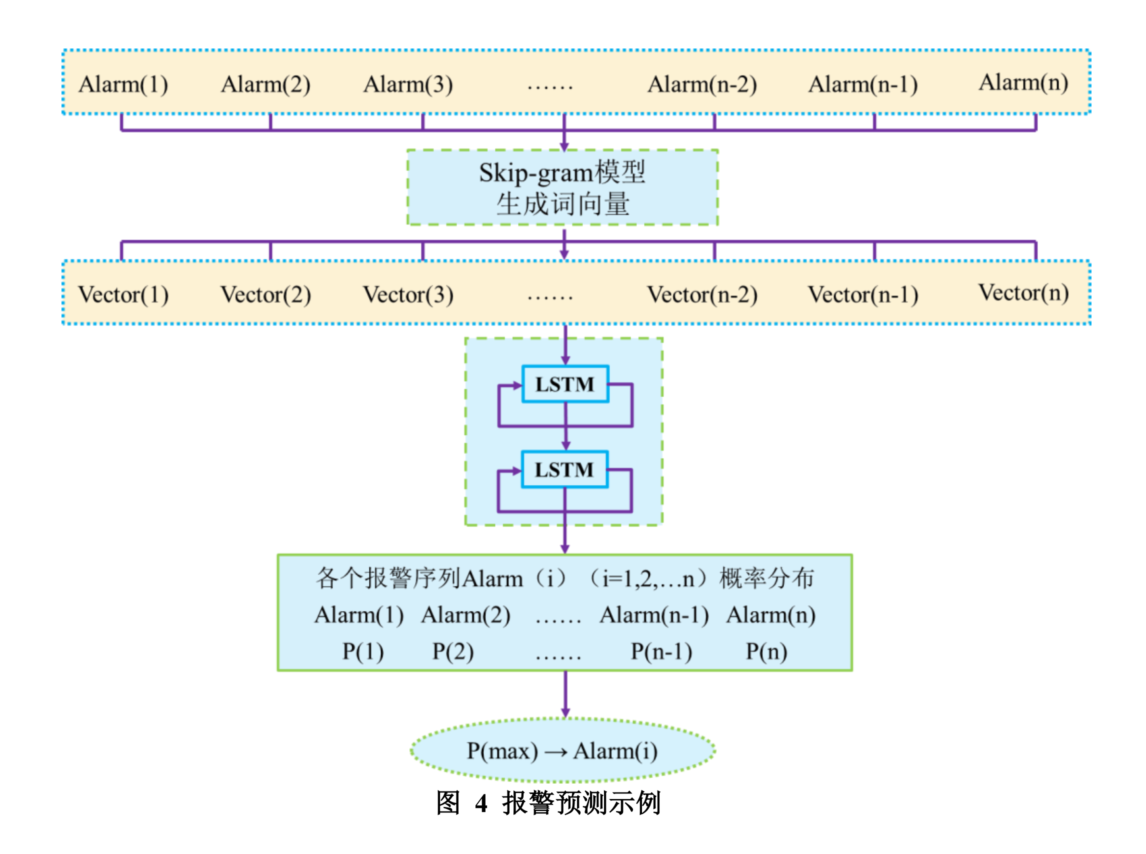



图 4 报警预测示例

# 4 案例分析

以某柴油加氢装置报警日志作为分析案例，运用 LSTM 报警预测方法。

## 4.1 报警日志预处理

文本清洗（将报警日志中无意义的数字或字母、非文本数据等影响分类效果或降低运算速度的不利因素去除，如消除冗余抖振报警）。表 1 为某柴油加氢装置报警日志的部分原始文本数据，它提供了报警名称、报警等级、报警描述、时间戳等有用信息，同时里面也包含一些无用信息，如数据区信息等。

表 1 某柴油加氢装置报警日志原始文本数据表

| 序号 | 报警时间 | 数据区 | 位号名 | 报警描述 | 内容 | 优先级 |
|---|---|---|---|---|---|---|
| 1 | 13-01-25 12:00:43 | 0 区 | TI7402_2B | P1001B 轴承温度 | HI | 0 |
| 2 | 13-01-25 11:37:52 | 0 区 | TI7402B | P1001B 轴承温度 | HI | 0 |
| 3 | 13-01-25 04:19:49 | 0 区 | FIQ3315.VAL | 循环水流量 | L0 | 0 |
| 4 | 13-01-25 11:09:28 | 0 区 | TRC3203 | 空冷 A3002 出口温控 | L0 | 0 |
| 5 | 13-01-07 21:22:38 | 0 区 | TRC3105 | 空冷 A3001 出口温控 | L0 | 0 |
| 6 | 13-01-24 22:36:22 | 0 区 | PIA3053D | C1002A 供油总管压力 | L0 | 0 |
| 7 | 13-01-24 22:30:29 | 0 区 | TI4053 | 瓦斯气温度 | LL | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 8 | 13-01-24 22:30:28 | 0 区 | TI4053 | 瓦斯气温度 | L0 | 0 |
| 9 | 13-01-24 22:01:21 | 0 区 | TI3305 | 非净化压缩气进装置温度 | LL | 0 |
| 10 | 13-01-24 22:01:19 | 0 区 | TI3305 | 非净化压缩气进装置温度 | L0 | 0 |
| 11 | 13-01-24 15:28:44 | 0 区 | PI3117.VAL | P1001C 出口流量 | HI | 0 |
| 12 | 13-01-24 15:27:37 | 0 区 | PI3117.VAL | P1001C 出口流量 | HH | 0 |
| 13 | 13-01-24 15:14:44 | 0 区 | FR3105 | 氢气流量 | L0 | 0 |
| 14 | 13-01-24 15:08:19 | 0 区 | FRC3101.VAL | 泵 P1001A,B 出口流控 | HI | 0 |
| 15 | 13-01-24 14:54:48 | 0 区 | LICA3203 | 罐 V1007 界控 | L0 | 0 |
| 16 | 13-01-24 14:45:20 | 0 区 | FI3117.VAL | P1001C 出口流量 | L0 | 0 |
| 17 | 13-01-24 14:18:49 | 0 区 | LICA3203 | 罐 V1007 界控 | HI | 0 |
| 18 | 13-01-24 13:47:58 | 0 区 | TRC3208 | 换 E1004 出口温控 | HI | 0 |
| 19 | 13-01-24 10:44:47 | 0 区 | FI13117.VAL | P1001C 出口流量 | LL | 0 |

表 2 为某柴油加氢装置报警日志的进行清洗处理后的结果，可以看出清洗过后信息更为简洁集中，清洗后的数据对应了知识图谱中的若干三元组，将报警描述转换为英文在模型训练时表现出更好的算力，直接调用 NLTK 即可，不需要通过 Jieba 分词工具调用停用词词典和用户自定义词典。

**表 2 某柴油加氢装置报警日志数据清洗表**

| 序号 | 时间戳 | 位号 | 具体设备 | 涉及因素 | 等级 |
|---|---|---|---|---|---|
| 1 | 2013/1/25 12:00 | TI7402B | P1001B | temperature | HI |
| 2 | 2013/1/25 4:19 | FIQ3315.VAL | circulating water | flow | L0 |
| 3 | 2013/1/25 11:09 | TRC3203 | A3002 | temperature control | L0 |
| 4 | 2013/1/7 21:22 | TRC3105 | A3001 | temperature control | L0 |
| 5 | 2013/1/24 22:36 | PIA3053D | C1002A | pressure | L0 |
| 6 | 2013/1/24 22:30 | TI4053 | gas | temperature | LL |
| 7 | 2013/1/24 22:30 | TI4053 | gas | temperature | L0 |
| 8 | 2013/1/24 22:01 | TI3305 | unpurified compressed gas | temperature | LL |
| 9 | 2013/1/24 22:01 | TI3305 | unpurified compressed gas | temperature | L0 |
| 10 | 2013/1/24 15:28 | PI3117.VAL | P1001C | flow | HI |

4.2 词嵌入建模

使用 Skip-gram 模型来学习每个报警变量的向量表示，可以捕捉到报警变量之间的语义和语法关系。使用 Gensim 这个高效的 Python 库来训练 Skip-gram 模型，它需要把每个报警序列作为输入，每个序列由按时间顺序排列的报警变量组成。Skip-gram 模型会自动地将每个报警变量映射到一个 $n$ 维的向量空间中，从而得到一个嵌入矩阵，大小为 $s \times n$ ，其中 $s$ 报

警变量类型的个数，$n$ 为嵌入空间的维度。此处将报警日志数据变量通过向量化变为 123 个 80 维的向量。Skip-gram 模型训练过程参数见表 2。

**表 3 Skip-gram 模型参数**

| 参数名 | 参数值 |
|---|---|
| Vector_size | 80 |
| windows | 4 |
| Epoch | 100 |
| Min_count | 1 |
| Workers | 10 |
| Negative | 30 |

4.3LSTM 预测

TensorFlow 是一个开源的软件库，它可以利用数据流图来进行高效的数值计算。数据流图是一种有向图，其中每个节点表示一个数学运算，每条边表示一个多维数组。TensorFlow 提供了一个灵活的 API，可以让用户根据需要，将计算任务分配到不同的设备上执行，例如桌面、服务器或移动设备中的 CPU 或 GPU。这样可以提高计算效率和并行性。TensorFlow 在机器学习或深度学习的网络研究中有着广泛的应用，它可以支持多种类型的网络结构和算法。

本文使用 Python 语言和 TensorFlow 软件包来构建长短期记忆网络模型，该模型是一种循环神经网络，它可以处理时序数据，并具有长期记忆能力。长短期记忆网络模型要求输入数据是实数或者数字化数据，因此本文使用报警向量作为训练数据。报警向量是通过 Skip-gram 模型从报警序列中学习得到的，它可以反映报警变量之间的语义和语法关系。本文设定长短期记忆网络模型的步长为 ，即以 个报警向量作为输入，预测第 个报警向量。这样可以实现对报警序列的预测和分析。本文还对长短期记忆网络模型的其他参数，如向量嵌入的维度、训练批次量等进行了合理的设置和调整，以提高模型的性能和准确度，得到了不错的模型。具体参数见表 3。

**表 3 长短期记忆网络模型参数**

| 参数名 | 参数值 |
|---|---|
| Stepstime | n |
| Optimezer | Adam |
| Loss | Mae/MSe |

| | |
|---|---|
| Epoch | 100 |
| Batch_size | 72 |
| LSTM 遗忘设置 | 0.1 |
| Dropout 概率 | 0.2 |

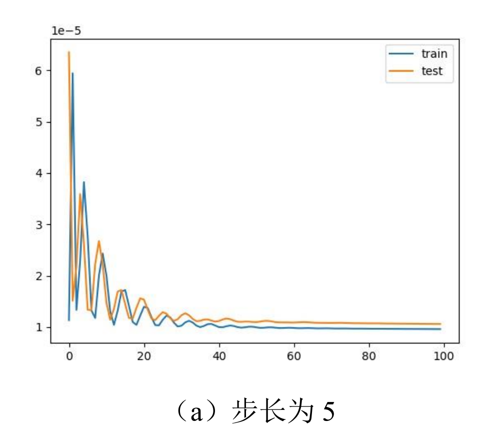
长短期记忆网络模型由两层组成，第一层是一个循环神经网络层，可以处理输入的时间序列数据，并输出 80 维的向量。这个层使用了 Dropout 来防止过拟合，也使用了 bias 来增强记忆能力。第二层是一个全连接层，可以将 80 维的向量映射到一个标量，作为预测值。这个层使用了线性激活函数，以保持输出的连续性。模型使用平均绝对误差作为损失函数，使用 Adam 作为优化器。平均绝对误差可以衡量预测值和真实值之间的差距，Adam 可以自适应地调整学习率和动量，以加速收敛过程。

对比不同步长下的预测结果如下：

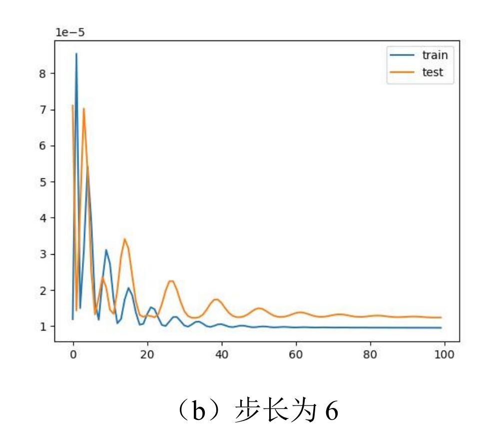
（1）损失函数选择 mse，步长 n=5、6、7、8。因为数据较少的原因选择训练轮次为 100，不同步长下的损失对比见表 4。可以看出训练集的损失稳定在 10-6 量级，验证集的损失稳定在 10-5 量级；随着步长的增加训练集和验证集的损失均有下幅度下降，但很小；图 5 展示了不同步长下的损失曲线，发现曲线在训练轮次很小的时候就收敛了，之后出现了周期性的波动，但是总体 loss 呈现下降趋势并趋于稳定。

**表 4 mse 下不同步长损失对比**

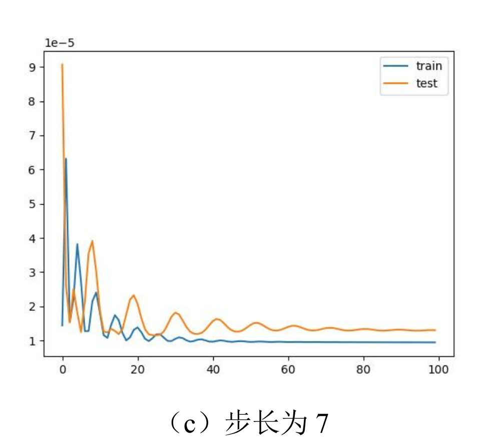
| 步长 | Loss | Val_loss |
|---|---|---|
| 5 | 9.5781e-06 | 1.0571e-05 |
| 6 | 9.5548e-06 | 1.2415e-05 |
| 7 | 9.5548e-06 | 1.2415e-05 |
| 8 | 9.3899e-06 | 1.0478e-05 |



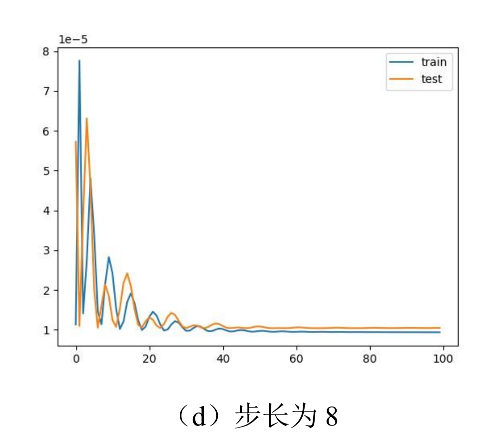

（a）步长为 5




（b）步长为 6




（c）步长为 7



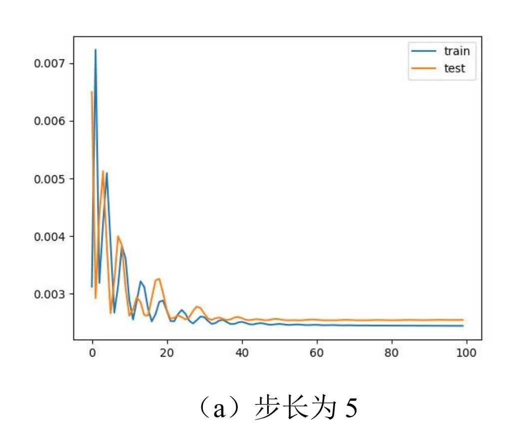

（d）步长为 8

**图 5 mse 下不同步长对比**

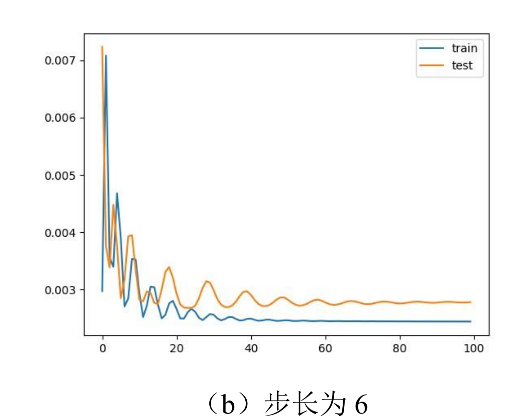
（2）损失函数选择 ame，步长 n=5、6、7、8.不同步长下的损失对比见表 5。可以看出训练集的损失稳定在 0.0024 左右，验证集的损失稳定在 0.0026 左右；随着步长的增加训练集和验证集的损失均有下幅度下降，但很小；图 6 展示了不同步长下的损失曲线，发现曲线在训练轮次很小的时候就收敛了，之后出现了周期性的波动，但是总体 loss 呈现下降趋势并趋于稳定。

**表 5 mae 下不同步长损失对比**

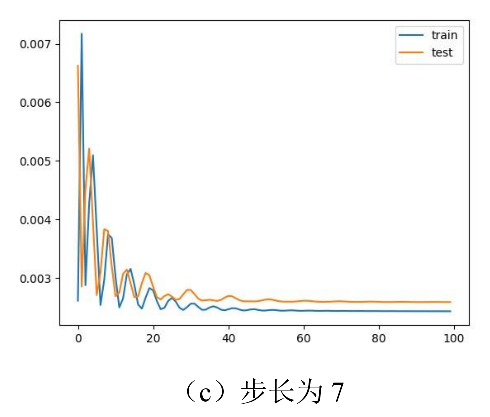
| 步长 | Loss | Val_loss |
|---|---|---|
| 5 | 0.0024 | 0.0025 |
| 6 | 0.0024 | 0.0028 |
| 7 | 0.0024 | 0.0026 |
| 8 | 0.0024 | 0.0026 |



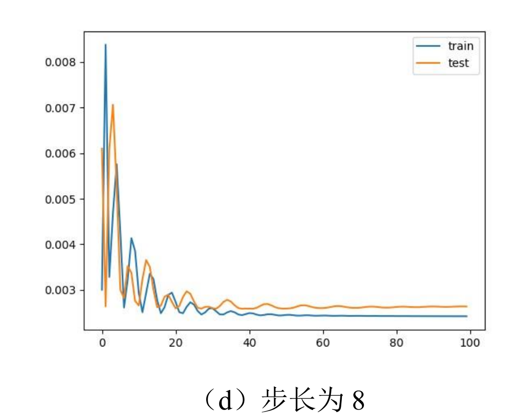

（a）步长为 5




（b）步长为 6




（c）步长为 7




（d）步长为 8

**图 6 mae 下不同步长对比**

得到训练完成的长短期记忆网络模型（将表 2 数据的 2/3 作为训练集，1/3 作为输入进行预测），将数据导入长短期记忆网络模型进行预测得到如下所示的预测报警词向量，准确率为 85.6%

[-1.3352952e-03，-1.9064530e-03，-5.2710669e-04，-3.7935303e-04，-5.3319125e-04，-4.5719827e-04，-1.0022228e-03，3.2839796e-04，4.3558679e-04，8.1800204e-04，2.8444192e-04，1.1573574e-03，1.6119785e-03，2.3732153e-03，2.5206394e-03，2.1843205e-03，2.5808737e-03，2.5722622e-03，2.9517047e-03，2.6353092e-03，1.2544806e-03，1.2108956e-03，1.1859352e-03，1.3071689e-03，1.3748782e-03，1.4311392e-03，1.2827201e-03，1.8794824e-03，1.8889243e-03，2.3909144e-03，2.7460307e-03，2.3870482e-03]

4.4 可视化处理

在得到的预测结果的词向量的基础上，调用保存的词向量模型，反向搜索与其相似度最高的文本词，确定下一个报警最可能发生的位置，即得到位号。5.4.3 节得到的预测结果对应的位号如图 7 所示。“FRA3101”的概率为 0.2875，“FRC3212”的概率为 0.2717，二者相差不大。从安全的角度考虑，将二者都作为预测结果返回报警日志进行搜索。

```
FRA3101 0.28753817081451416
FRC3212 0.27169206738471985
```

**图 7 预测结果示例**

如图 8 所示，通过位号确定可能发生的具体报警描述，在知识图谱上进行可视化。内容包括了位号、具体设备、涉及因素、报警等级四个要素。当输入不同的、更大体量的报警序

列时，这种对比将更加突出，通过对应色块数量的直观数量感受可以很快的分辨出危险度较高的设备节点，从而达到优化报警管理的目的。

## 5 结论

本文针对报警预测这一重要的工业应用问题，提出了一种基于长短时记忆网络的报警预测方法。利用知识图谱的结构和语义信息，对报警序列进行了丰富和标准化的表示。然后，将报警序列中的每个词通过 Word2Vec 嵌入算法转换为低维稠密向量，作为长短期记忆网络模型的输入。

长短期记忆网络模型是一种能够处理时序数据的循环神经网络，它能够学习报警序列中的长短期依赖关系，并根据当前的输入和历史状态，预测下一个可能发生的报警。为了提高预测性能，并通过实验对比了不同步长下损失函数的收敛速度和波动情况，选择了最优的步长参数。本文通过在真实的工业数据集上进行了大量的实验，验证了该方法在报警预测任务上的有效性和优越性，为故障诊断和预防提供了有价值的参考。将预测所得的结果加入知识图谱中，完成了知识图谱的演化，并将预测节区别于原本的节点，使得知识图谱结构更加完善。

长短期记忆网络模型比较单一，在其基础上演化出 BiLSTM 等一众衍生模型，通过更优模型可增加模型预测准确率和预测的效果，这样知识图谱的演化效果更佳。